EMbaRC

European Consortium of Microbial Resource Centres

Grant agreement number: **228310**

Seventh Framework Programme

Capacities

Research Infrastructures

Combination of Collaborative Project and Coordination and Support Actions

Deliverable D.15.20 formerly: D.JRA2.2.1

Title: **New Database containing sequences for eukaryote strains and species assessed for their authentication**

Due date of deliverable: M29

Actual date of submission: M32

Start date of the project: 1st February 2009 **Duration:** 44 months

Organisation name of the lead beneficiary: INRA

Version of this document: V1.0

Dissemination level: PU

PU	**Public**	**X**
PP	Restricted to **other programme participants** (including the Commission)	
RE	Restricted to a **group defined by the Consortium** (including the Commission)	

EMbaRC is financially supported by the Seventh Framework Programme (2007-2013) of the European Communities, Research Infrastructures action

Document properties	
Project	EMbaRC
Workpackage	JRA2
Deliverable	D.15.20, formerly: D.JRA2.2.1
Title	New Database containing sequences for eukaryote strains and species assessed for their authentication
Version number	V1.0
Authors	WEISS Stéphanie (INRA)
Abstract	A database which contains carefully checked sequences from all Saccharomycotina type strains has been constructed. Tools for sequence comparison has been implemented in the database which allow to compare query sequences to the sequences in the database in order to provide (i) identification at the species level (ii) robust phylogeny of major and well defined clades.
Validation process	Document prepared by INRA and submitted to the Executive Committee for agreement.

Revision table			
Date	***Version***	***Revised by***	***Main changes***
29/08/2011	0.1	Stéphanie Weiss (INRA)	Creation
02/09/2011	0.2	Christiane Bouchier (IP, WP leader)	Validation
16/09/2011	0.3	Sylvie Lortal (INRA, Coordinator)	Significance appended
21/09/2011	0.4	Yohan Lecuona (INRA, Project Manager)	Layout adjustments
06/10/2011	1.0	""	Final version for submission

Contents

Background and Objectives

The taxonomy of ascomycetous yeasts has greatly evolved in the genomic era. Genomic studies yielded large amount of sequences, which were subsequently used to improve yeast phylogeny with multi-genic analyses. This led and is still leading to frequent species name changes. Currently, information on yeast taxonomy associated to sequences can be found in various databases, which are not up-to-date or/and inexact. Finally, sequences of interest for taxonomy are diluted in environmental and sequences used for typing are difficult to retrieve from generalist databases. Within the frame of the JRA2.2.1 task, the construction of a relevant database with verified nucleotidic sequences and an interface allowing the users to obtain correct identification, taxonomy and phylogeny of yeasts species, was undertaken.

The aim of this task is to screen and validate the most frequently targeted sequences for microbial systematics of yeasts, i.e. ribosomal RNA sequences (26S D1/D2, intergenic transcribed sequences ITS...), sequences encoding actin, elongation factor, mitochondrial genes...

Sequences from a large number of ascomycetous yeast species will be retrieved from DNA databases. Quality of sequence will be assessed. A database which contains carefully checked sequences from a number of type strains will be constructed. Several tools of sequence comparison (blast, fasta etc…) will be implemented in the database which will allow comparison of provided sequences to the sequences in the database in order to provide:

- identification at the species level
- classification within well defined clades or subspecies

Efficient data-processing bioinformatics programs (Figenix, Phylogeny.fr) will be used to integrate the emerging data for the design of comprehensive phylogenetic trees. Ultimately, these databases will evolve towards a web service for EMbaRC and will enable to revisit the phylogeny of the relevant groups of microorganisms. It will thus represent an essential tool for the study of phylogeny-function relations.

1 Preliminary work

The FunGeneDB database (http://www.fungene-db.org), developed at INRA CIRM-CF (Marseille, France), comprises well annotated and curated ribosomal DNA sequences in order to assist reliable identification of Basidiomycetes Polyporales (1). For the development of the YeastIP database, INRA CIRM-Levures (Grignon, France) received a skill transfer from CIRM-CF. FunGeneDB has been imported to the Topaze server (INRA Jouy-en-Josas, France), and, first, the

script was adapted to a Linux environment. Then, the SQL database was adapted to accommodate yeasts. Multiple modifications and improvements have been introduced since these first steps.

2 Database structure and content

The database was constructed with the MySQL software. Before their introduction in the database, sequences from generalist databases like GenBank/EMBL were screened through an expert selection (Fig.1). On the first of August 2011, YeastIP contains 4183 sequences that were extracted from NCBI and checked for quality and taxonomy relevance by using the just published comprehensive book on yeast taxonomy (2). The sequences in the YeastIP database represent 60 clades, 82 genera and 906 species and a choice of up to 10 markers per species, when available. The YeastIP database is therefore exhaustive since all the described Saccharomycotina yeast species are represented in the database

2.1 Strains

Priority has been given to the type strains sequences, according to Kurtzman *et al.* (2). In some particular cases, i.e. a non type strain presenting more sequences than the type strain or no sequences being available for the type strain, sequences of a non type strain have been introduced.

2.2 Markers

The implemented markers were choosen according to recent multigene taxonomic study (3-9). Retained markers are : entire 26S rDNA gene, entire 18S rDNA gene, D1/D2 part of the 26S rDNA gene, ITS1-5.8S-ITS2 intergenic sequence, TEF1-α partial gene, ACT1 partial exon, RPB1 and RPB2 partial gene, mtSm mitochondrial ribosomal small subunit partial gene and COXII mitochondrial gene. Each sequence is unique in the database and all type strains are represented by at least their D1/D2 sequence.

2.3 Introduction of sequences in the database

The sequences were retrieved from NCBI with their accession numbers. Each sequence was carefully checked for quality (length, number of N, origin...), and some supplemental information were added or corrected, like species name, clade appartenance, species name synonyms and CBS collection number.

3 Interface

The interface for user was developped in HTML/PHP/JavaScript language. The interface was developed to propose a maximum of relevant information and choices to search the database. The database is online since the 24/06/2011 with the public URL:

http://genome.jouy.inra.fr/yeastip/index.php

3.1 Sequence file

A file containing all relevant information for each sequence is available i.e. current species name and synonyms of the species, origin of the sequence, link to the original publication and possible comments.

3.2 Search sequences in database

The Search sequence tool allows the search for sequences *via* taxonomy information (clade, genus or species name), or keyword (text, species name synonym, CBS strain number…).

3.3 Authentication tool

The authentication tool using the Blast program was implemented to allow users to compare their sequences of interest to the sequences in the YeastIP database.

3.4 Marker table

The Search tool and the Authentication tool converge to the display of a table showing the available markers for each selected species or group of species (Fig. 2). This table will guide the users for the choice of markers to sequence in order to perform the most thorough phylogenic analysis with the largest number of species.

3.5 Concatenation and phylogeny tool

Since yeast phylogeny reconstructions now require multi-genic analysis, a tool to concatenate various sequences for each strain selected in the marker table was devised. The concatenation file is retrievable in Fasta format. Users can also add their own sequences of interest to create a concatenated file containing both their own sequences and the closely related sequences in YeastIP database. This file can be used to reconstruct phylogeny through a direct link to one of the currently best phylogeny plate-form, Phylogeny.fr (http://www.phylogeny.fr).

3.6 Help

A large part of the website is devoted to the user assistance. The help page contains not only information about the use of the website, but also a support in phylogeny to help user to do the best choice for their analysis.

4 Sustainability and update of the database

An administrator interface has been implemented to the website (secure pages, not visible publicly) to facilitate sustainability of the database by the CIRM-Levures BRC (http://www.inra.fr/cirmlevures).

The YeastIP database will be updated in collaboration with the CBS (http://www.cbs.knaw.nl/). The CBS, which receives for deposit the type strains of all newly described species, will provide information, sequences or DNA to the YeastIP database.

Conclusion

YeastIP is a unique database for fungal species; it facilitates the retrieval of taxonomic information and guides users to obtain robust phylogenic analyses. This work is linked to another part of the EMbaRC project, consisting in producing new sequences, which will feed the YeastIP database. YeastIP put the emphasis on multigenic analysis to improve good practice in hemiascomycetous yeasts phylogeny, and could be extended to all fungal species. The promotion of this work will be ensured by the article in preparation: YeastIP: a database for identification and phylogeny of ascomycetous yeasts, Weiss S., Samson F., Navarro D., Casaregola S.

References

(1) **FunGeneDB: a bioinformatic tool as an aid to identification of Polyporales.** Navarro D., Favel A., Chabrol O., Haon M., Taussac S., Pontarotti P., Delattre M., Welti S. and Lesage-Meessen L. (submitted)

(2) **The Yeast, a Taxonomical study**, 5th edition, Elsevier Amsterdam. Kurtzman C.P., Boekhout T. and Fell J., 2011.

(3) **Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences.** Kurtzman C.P. and Robnett C.J., 1998, *Antonie van Leeuwenhoek* , 73:331-371.

(4) **Partial sequence analysis of the actin gene and its potential for studying the phylogeny of *Candida* species and their teleomorphs.** Daniel H.M., Sorell T.C. and Meyer W., 2001, *Int. J. Syst. Evol. Microbiol.*, 51:1593-1606.

(5) **Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analysis.** Kurtzman C.P. and Robnett C.J., 2003, *FEMS Yeast Res.*, 3:417-432.

(6) **Phylogenetic circumscription of *Saccharomyces, Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma* and *Zygotorulaspora*.** Kurztman C.P., 2003, *FEMS Yeast Res.*, 4:233-245.

(7) **Evaluation of ribosomal RNA and actin gene sequences for the identification of ascomycetous yeasts.** Daniel H.M. and Meyer W., 2003, *Int. J. Food Microbiol.*, 86:71-78.

(8) **Multigene phylogenetic analysis of *Trichomonascus, Wickerhamiella* and *Zygoascus* yeast clade, and the proposal of *Sugiyamaella* gen. nov. and fourteen new species combinations.** Kurtzman C.P. and Robnett C.J., 2007, *FEMS Yeast Res.*, 7:141-154.

(9) **Re-examining the phylogeny of clinically relevant *Candida* species and allied genera based on multigene analysis.** Tsui C.K.M., Daniel H.M., Robert V. and Meyer W., 2008, *FEMS Yeast Res.*, 8:651-659.

Annexes

Figure 1: Database construction.

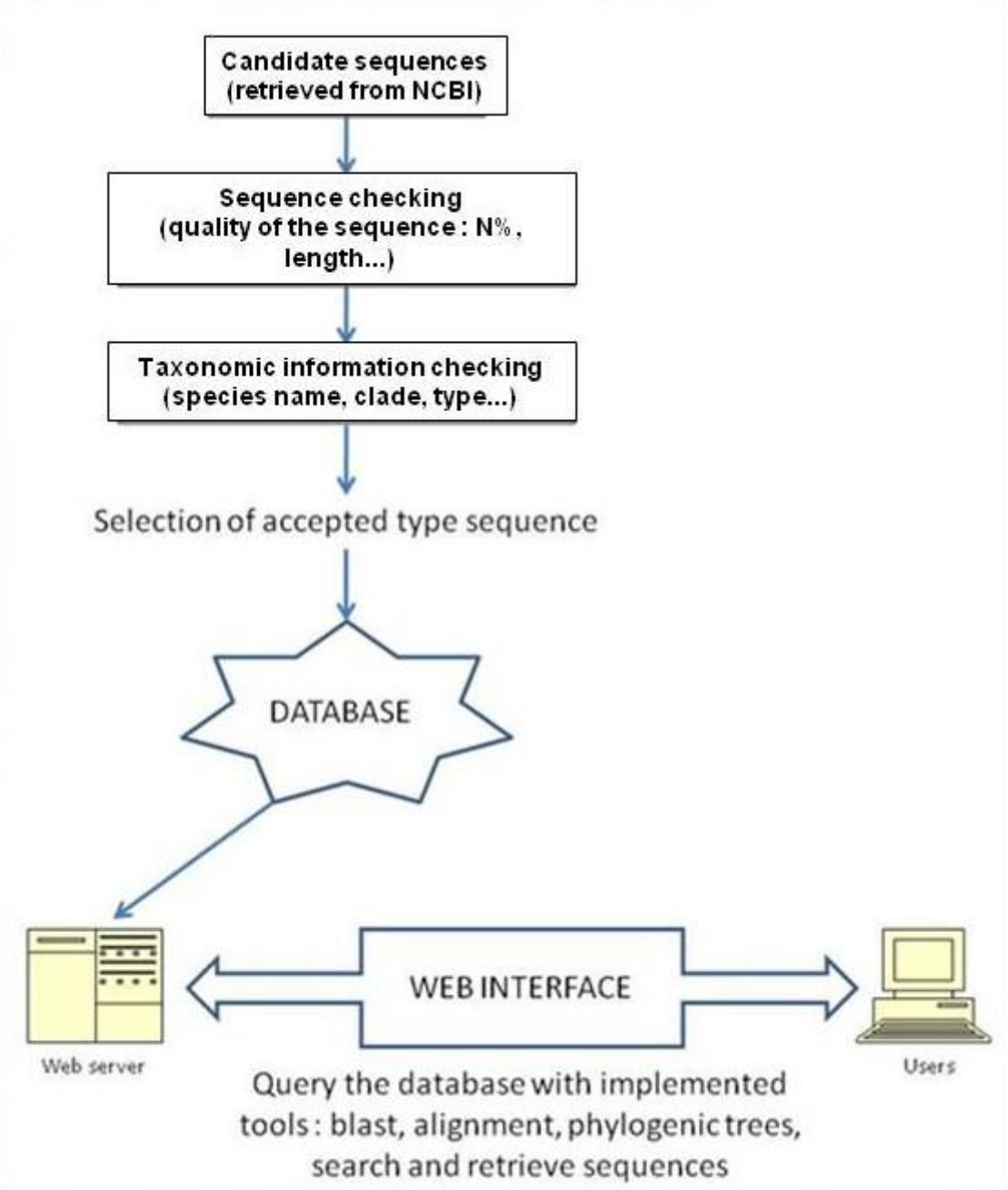
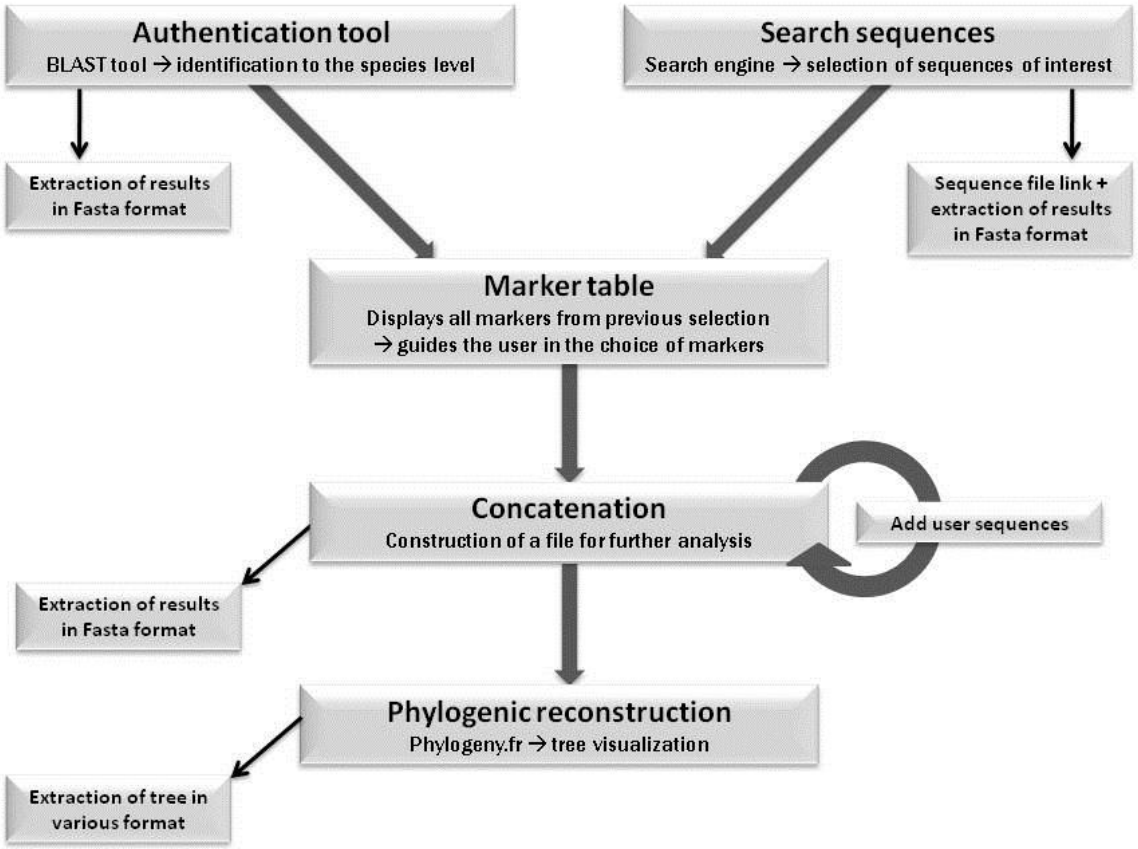


Figure 2: YeastIP interface processus.



Significance of this deliverable

The YeastIP database is a new unique and user-friendly tool for fungal taxonomy based on carefully selected sequences. It is also exhaustive for all yeast species described in the literature. Presently dedicated to hemiascomycetous yeasts, it can be extended to other fungal species.